1 Connection between VB and EM

The VB method we consider in the paper is closely related to the well known EM algorithm. According to TZIKAS et al. (2008), the principle of EM algorithm for the MAP estimation problem $\hat{\theta} = \arg\max_{\theta} p(\theta | \text{data})$ can be interpreted as follows: First of all, the log posterior distribution can be decomposed as

$$
\ln p(\theta | \text{data}) = \int q(z) \ln \frac{p(\theta, z | \text{data})}{q(z)} \, dz - \int q(z) \ln \frac{p(z | \theta, \text{data})}{q(z)} \, dz
\equiv L(q(z)) + \text{KL}(q(z)||p(z|\theta, \text{data}))
$$

(1.1)

where $z = \{z_1, ..., z_m\}$ are the hidden variables (to be integrated out), which are not of main interest, but are helpful to estimate the value of parameters $\theta$. $q(z)$ is a free selected distribution for $z$. Since $\text{KL}(q(z)||p(z|\theta, \text{data})) \geq 0$, $L(q(z))$ can be viewed as a lower bound of the log posterior $\ln p(\theta | \text{data})$. Denote the current value of $\theta$ as $\theta^{\text{OLD}}$. In the E-step, the lower bound is maximized with respect to the distribution $q(z)$, which leads to $q(z) = p(z|\theta^{\text{OLD}}, \text{data})$. In the M-step, the lower bound with $q(z)$ substituted by $p(z|\theta^{\text{OLD}}, \text{data})$ is maximized with respect to $\theta$, which leads to a new estimation of $\theta$ as $\theta^{\text{NEW}}$. By implementing these two steps successively, the value of log posterior function increases, and eventually reaches its maximum.

An easy implementation of the EM algorithm requires the full conditional distribution $p(z|\theta, \text{data})$ to be explicitly known. However, in many models, hidden variables may interact with each other, and $p(z|\theta, \text{data})$ becomes too complicated to be used in EM. One solution is to implement the variational approximation method in the E-step, that is, the distribution $q(z)$ is assumed to be of the factorized form $q(z) = \prod_{i=1}^{m} q(z_i)$. Next, the lower bound can be maximized with respect to $q(z_i) \ (i = 1, 2, ..., m)$. This procedure has been called variational EM (see BEAL 2003). Furthermore, it is possible to consider the parameters $\theta$ as a part of hidden variables, and implement a variational algorithm with only the E-step. This is exactly the same VB method introduced in the paper. Therefore, variational Bayes and variational EM can be viewed as extensions of
traditional EM algorithm, and can be applied to a broader class of Bayesian models.

2 VB estimation algorithm for Bayesian Adaptive Shrinkage

The priors are specified as

\[ p(\beta_0) \propto 1, \quad (2.1) \]

\[ p(\tau_0^2) \propto \frac{1}{\tau_0^2}, \quad (2.2) \]

\[ p(\beta_j | \tau_j^2) \propto N(\beta_j | 0, \frac{1}{\tau_j^2}), \quad (2.3) \]

\[ p(\tau_j^2) \propto \frac{1}{\tau_j^2}, \quad (2.4) \]

and the logarithm of the joint distribution of parameters and data is

\[ \ln p(\theta, y) = p(\beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \tau_1^2, \ldots, \tau_p^2 | y) \]

\[ = \left[ \frac{n}{2} \ln \tau_0^2 - \frac{\tau_0^2}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \right] + \left[ -\ln \tau_0^2 \right] + \left[ \frac{1}{2} \sum_{j=1}^{p} \ln \tau_j^2 - \frac{1}{2} \sum_{j=1}^{p} \tau_j^2 \beta_j^2 \right] \]

\[ + [\sum_{j=1}^{p} \ln \tau_j^2] + C, \quad (2.5) \]

where \( C \) is a constant. Similarly as in VB-EBL, we can derive the VB marginal distributions \( \hat{q}(\cdot | y) \) for parameters \( \beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \tau_1^2, \ldots, \tau_p^2 \) in VB-BAS:

(I) Derivation of \( \hat{q}(\beta_0 | y) \): \( \hat{q}(\beta_0 | y) \) is recognized as a normal distribution with mean

\[ E[\beta_0] = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} E[\beta_j] x_{ij}), \quad (2.6) \]

and variance

\[ \text{Var}[\beta_0] = \frac{1}{n E[\tau_0^2]}. \quad (2.7) \]

(II) Derivation of \( \hat{q}(\beta_j | y) \): \( q(\beta_j | y) \) is also a normal distribution, with mean

\[ E[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1} E[\tau_0^2] \sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{k \neq j} E[\beta_k] x_{ik}) x_{ij}, \quad (2.8) \]

and variance

\[ \text{Var}[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1}, \quad (2.9) \]
and in addition we need
\[ E[\beta_j^2] = E[\beta_j]^2 + \text{Var}[\beta_j]. \] (2.10)

(III) Derivation of \( \hat{q}(\tau_0^2|y) \): \( \hat{q}(\tau_0^2|y) \) is a gamma distribution with parameters
\[ a_1 = \frac{n}{2}, \] (2.11)
\[ b_1 = \frac{\sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{j=1}^{p} E[\beta_j]x_{ij})^2 + \text{Var}[\beta_0] + \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2 \text{Var}[\beta_j]}{2}, \] (2.12)
and the expectation is
\[ E[\tau_0^2] = \frac{a_1}{b_1}. \] (2.13)

(IV) Derivation of \( \hat{q}(\tau_j^2|y) \): \( \hat{q}(\tau_j^2|y) \) is a gamma distribution with parameters
\[ a_{2j} = \frac{1}{2}, \] (2.14)
\[ b_{2j} = \frac{E[\beta_j^2]}{2}, \] (2.15)
and the expectation is
\[ E[\tau_j^2] = \frac{a_{2j}}{b_{2j}}. \] (2.16)

Additionally, we initialize the expectation \( E[\beta_j]|(j = 0, ..., p) \) as 0, and \( E[\tau_j^2]|(j = 0, ..., p) \) as 2, and then the approximate distributions in (I)-(IV) can be updated successively. Finally, we can derive the lower bound as
\[
L = -\frac{n - 1}{2} \ln(2\pi) + \frac{p + 1}{2} + \ln \text{Var}[\beta_0] + \sum_{j=1}^{p} \frac{\ln \text{Var}[\beta_j]}{2} - a_1 \ln b_1 + \ln \Gamma(a_1) \\
- \sum_{j=1}^{p} a_{2j} \ln b_{2j} + \sum_{j=1}^{p} \ln \Gamma(a_{2j}), \tag{2.17}
\]
where \( \Gamma(\bullet) \) is the gamma function.

3 VB estimation algorithm for Bayesian LASSO

The priors are specified as
\[ p(\beta_0) \propto 1, \] (3.1)
\[ p(\tau_0^2) \propto 1/\tau_0^2, \] (3.2)
\[ \tau_j^2 \propto \mathcal{N}(\beta_j|0, \frac{1}{\tau_j^2}), \] (3.3)
\[ p(\tau_j^2 | \lambda^2) \propto \text{Inv-Gamma}(\tau_j^2 | 1, \frac{\lambda^2}{2}), \quad (3.4) \]
\[ p(\lambda^2 | \gamma, \nu) \propto \text{Gamma}(\lambda^2 | \gamma, \nu), \quad (3.5) \]

and the logarithm of the joint distribution of parameters and data is

\[
\ln p(\theta, y) = p(\beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \tau_1^2, \ldots, \tau_p^2, \lambda^2 | y)
\]
\[
= \left[ \frac{n}{2} \ln \tau_0^2 - \frac{\tau_0^2}{2} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \right] + \left[ \frac{1}{2} \sum_{j=1}^{p} \ln \tau_j^2 - \frac{1}{2} \sum_{j=1}^{p} \tau_j^2 \beta_j^2 \right]
\]
\[
+ \left[ \frac{n}{2} \ln \lambda^2 - 2 \sum_{j=1}^{p} \ln \tau_j^2 - \frac{n}{2} \lambda^2 \right] + \left[ (\gamma - 1) \ln \lambda^2 - \nu \lambda^2 \right] + C, \quad (3.6)
\]

where \( C \) is a constant, and a VB algorithm is derived as following:

(I) Derivation of \( \hat{q}(\beta_0 | y) \): \( \hat{q}(\beta_0 | y) \) is recognized as a normal distribution with mean

\[ E[\beta_0] = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} E[\beta_j] x_{ij}), \quad (3.7) \]

and variance

\[ \text{Var}[\beta_0] = \frac{1}{n} \frac{1}{E[\tau_0^2]}, \quad (3.8) \]

(II) Derivation of \( \hat{q}(\beta_j | y) \): \( \hat{q}(\beta_j | y) \) is also a normal distribution with mean

\[ E[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1} E[\tau_0^2] \sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{k \neq j} E[\beta_k] x_{ik}) x_{ij}, \quad (3.9) \]

and variance

\[ \text{Var}[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1}, \quad (3.10) \]

and in addition we need

\[ E[\beta_j^2] = E[\beta_j]^2 + \text{Var}[\beta_j]. \quad (3.11) \]

(III) Derivation of \( \hat{q}(\tau_0^2 | y) \): \( \hat{q}(\tau_0^2 | y) \) is a gamma distribution with parameters

\[ a_1 = \frac{n}{2} \quad (3.12) \]
\[ b_1 = \sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{j=1}^{p} E[\beta_j] x_{ij})^2 + \text{Var}[\beta_0] + \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2 \text{Var}[\beta_j], \quad (3.13) \]

and the expectation is

\[ E[\tau_0^2] = \frac{a_1}{b_1}. \quad (3.14) \]
(IV) Derivation of \( \hat{q}(\tau_j^2 | y) \): \( \hat{q}(\tau_j^2 | y) \) is an inverse Gaussian distribution with parameters

\[
\mu = \sqrt{\frac{E[\lambda^2]}{E[\beta_j^2]}} \quad (3.15)
\]

and

\[
\lambda_0 = E[\lambda^2], \quad (3.16)
\]

and in addition we need the expectation of \( \tau_j^2 \) as

\[
E[\tau_j^2] = \mu, \quad (3.17)
\]

and the expectation of \( \frac{1}{\tau_j^2} \) as

\[
E\left[ \frac{1}{\tau_j^2} \right] = \frac{1}{\mu} + \frac{1}{\lambda_0}. \quad (3.18)
\]

(V) Derivation of \( \hat{q}(\lambda^2 | y) \): \( \hat{q}(\lambda^2 | y) \) is a gamma distribution with parameters

\[
a_2 = p + \gamma, \quad (3.19)
\]

and

\[
b_2 = \frac{1}{2} \sum_{j=1}^{p} E\left[ \frac{1}{\tau_j^2} \right] + v, \quad (3.20)
\]

and the expectation of \( \lambda^2 \)

\[
E[\lambda^2] = \frac{a_2}{b_2}. \quad (3.21)
\]

Before running iterative algorithm, we may initialize the expectation \( E[\beta_j] (j = 0, \ldots, p) \) as 0, \( E[\tau_j^2] (j = 0, \ldots, p) \) as 2, and \( E[\lambda^2] \) as 1. Finally, the lower bound is

\[
L = -\frac{n - p - 1}{2} \ln(2\pi) + \frac{2p + 1}{2} - p \ln 2 + \sum_{j=1}^{p} E[\beta_j^2] E[\tau_j^2] + \frac{\ln \text{Var}[\beta_0]}{2} + \sum_{j=1}^{p} \frac{\ln \text{Var}[\beta_j]}{2} - a_1 \ln b_1 + \Gamma(a_1) - \frac{a_1 \ln \lambda_0}{2} - a_2 \ln b_2 + \Gamma(a_2). \quad (3.22)
\]

4 MCMC algorithm for Extended Bayesian LASSO

The priors in EBL are specified as

\[
p(\beta_0) \propto 1, \quad (4.1)
\]

\[
p(\sigma_0^2) \propto 1/\sigma_0^2, \quad (4.2)
\]

\[
p(\beta_j | \sigma_j^2) \propto N(\beta_j | 0, \sigma_j^2), \quad (4.3)
\]
\[ p(\sigma_j^2|\delta^2, \eta_j^2) \propto \text{Exp}(\sigma_j^2 \frac{\delta^2 \eta_j^2}{2}), \tag{4.4} \]

\[ p(\delta^2|\gamma, \upsilon) \propto \text{Gamma}(\delta^2|\gamma, \upsilon), \tag{4.5} \]

and

\[ p(\eta_j^2|\psi, \vartheta) \propto \text{Gamma}(\eta_j^2|\psi, \vartheta). \tag{4.6} \]

A description of the differences between our approach and MUTSHINDA and SILLANPÄÄ (2010) can be seen in the Discussion section. A Gibbs sampler can be built based on the following full conditional posterior distributions for parameters \( \beta_0, \beta_1, \ldots, \beta_p, \sigma_0^2, \sigma_1^2, \ldots, \sigma_p^2, \delta^2, \eta_1^2, \ldots, \eta_p^2 \):

(I) Full conditional posterior distribution of \( \beta_0 \):

\[ p(\beta_0|\theta, \beta_0, y) \propto N(\beta_0 \mid \frac{1}{n} \sum_{j=1}^{n} (y_i - \sum_{j} x_{ij} \beta_j), \frac{1}{n} \sigma_0^2). \tag{4.7} \]

(II) Full conditional posterior distribution of \( \beta_j \):

\[ p(\beta_j|\theta, \beta_j, y) \propto N(\beta_j \mid \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{k \neq j} x_{ik} \beta_k)}{(\sum_{i=1}^{n} x_{ij}^2 + \frac{\sigma_j^2}{\sigma_0^2})}, \frac{\sigma_0^2}{\sum_{i=1}^{n} x_{ij}^2 + \sigma_j^2}). \tag{4.8} \]

(III) Full conditional posterior distribution of \( \sigma_0^2 \):

\[ p(\sigma_0^2|\theta, \sigma_0^2, y) \propto \text{Inv-Gamma}(\sigma_0^2 \mid \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j} x_{ij} \beta_j)^2). \tag{4.9} \]

(IV) Full conditional posterior distribution of \( \frac{1}{\sigma_j^2} \):

\[ p\left( \frac{1}{\sigma_j^2}|\theta, \sigma_j^2, y \right) \propto \text{Inv-Gaussian}(\frac{1}{\sigma_j^2} \mid \sqrt{\frac{\delta^2 \eta_j^2}{b_j^2}}, \delta^2 \eta_j^2). \tag{4.10} \]

(V) Full conditional posterior distribution of \( \delta^2 \):

\[ p(\delta^2|\theta, \delta^2, y) \propto \text{Gamma}(\delta^2|\gamma, \frac{1}{2} \sum_{j=1}^{p} \eta_j^2 \sigma_j^2 + \upsilon). \tag{4.11} \]

(VI) Full conditional posterior distribution of \( \eta_j^2 \):

\[ p(\eta_j^2|\theta, \eta_j^2, y) \propto \text{Gamma}(\eta_j^2|1 + \psi, \frac{1}{2} \delta^2 \sigma_j^2 + \vartheta). \tag{4.12} \]

In practice, to avoid underflow/overflow, we restrict the value of \( \sigma_j^2 (j = 1, \ldots, p) \) to the interval \([10^{-20}, 10^{20}]\).
Finally, descriptions of the Gibbs sampler for the BAS model can be found in XU (2003), and that for the BL model can be found in YI and XU (2008).

**Literature cited**


MUTSHINDA, C. M., and M. J. SILLANPÄÄ 2010 Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186: 1067–1075

